1

Introduction

1.1 EXCITEMENT AT THE INTERFACE OF COMPUTING AND BIOLOGY

Sustained progress across all areas of science and technology over the last half-century has transformed the expectations of society in many ways. Yet, even in this context of extraordinary advances, both the biological sciences and the computer and information sciences share a number of characteristics that are compelling.

First, both fields have been characterized by exponential growth, with doubling times on the order of 1-2 years. In information technology (IT), both the component density of microprocessors and the information storage density on hard disk drives have increased exponentially with doubling times from 9 to 18 months. In biology, the rate of growth of the biological literature is characterized by exponential growth as well (e.g., the growth in GenBank is on the order of 60 percent per year, a rate comparable to Moore's law for microprocessors). While these growth rates cannot continue indefinitely, exponential growth is likely at least in the short term.

Second, both fields deal with organisms and phenomena or artifacts of astounding complexity. Both biological organisms and sophisticated computer systems involve very large numbers of components and interconnections between them, and out of these assemblages of components and connections emerges interesting and useful functionality. In the information technology context, the significance of these connections and components is much better understood than in the biological context, not least because human beings have been responsible for the design of information technology systems such as operating systems and computer systems. Still, the capabilities of existing computing methodologies to design or characterize large-scale information systems and networks are being stretched, and in the biological domain, a systems-level understanding of biological or computer networks is both highly important and difficult to achieve. In addition, information technology is a necessary and enabling technology for the study of complex objects. Computers are the scientific instruments that let us see genomes just as electron microscopes let us see viruses, or radio telescopes let us see quasars.

Third, both biology and information technology have profound and revolutionary implications for science and society. From an intellectual standpoint, biology offers at least partial answers to eternal questions such as, What is life? Also, biological science and technology have the potential for great impact on human health and well-being, including improved disease treatments, rapid environmental cleanup, and more robust food production. Computing and information technology enable human beings to acquire, store, process, and interpret enormous amounts of information, and continue to underpin much of modern society.

Finally, several important areas of interaction between the two fields have already emerged, and there is every expectation that more will emerge in the future. Indeed, the belief of the committee that there are many more synergies at the interface between these two fields than have been exploited to date is the motivation for this report. Against this backdrop, it makes good sense to consider potential interactions between the two fields—what this report calls the "BioComp" interface.

As for the nature of computing that can usefully be exploited by life scientists, there is a range of possibilities. For some problems encountered by biology researchers, a very rudimentary knowledge of computing and information technology is quite sufficient. However, as problems become bigger and/or more complex, what one may pick up by hacking and reading manuals is no longer sufficient. To address such problems, the kinds and levels of expertise needed are more likely to require significant formal study of computer science (e.g., as an undergraduate major in the field). And for still more difficult, larger, or more complex problems, the kinds and levels of expertise needed stretch the current state of knowledge of the field—a point that illuminates the importance of real computer science research in a biological context.

Nor is the utility of computing limited to providing tools or models—no matter how sophisticated—for biologists to use. As discussed in Chapter 6, computing can also provide intellectual abstractions that may provide insight into biological phenomena and a useful language for describing such phenomena. As one example, notions of circuit and network and modularity—originally conceptualized in the world of engineering and computer science—have much applicability to understanding biological phenomena.

On the other side, biology refers to the scientific study of the activities, processes, mechanisms, and other attributes of living organisms. For the purposes of this report, biology, biomedicine, life sciences, and other descriptions of research into how living systems work should be regarded as synonymous. In this context, for the past decade, researchers have spoken increasingly of a new biology, a biology of the 21st century, one that is driven by new technologies, that is more automated with tools and methods provided by industrial models, and that often entails high-throughput data acquisition.[1] This report examines the BioComp interface from the perspective of 21st century biology, as a science that integrates traditional empirical and experimental biology with a systems-level biology that considers the multiscale, hierarchical, highly interwoven, or interactive aspects intrinsic to living systems.

1.2 PERSPECTIVES ON THE BIOCOMP INTERFACE

This report addresses computationally inspired ways of understanding biology and biologically inspired ways of understanding computing. Although the committee started its work with the idea that it would discover a single community and intellectual synthesis of biology and computing, closer examination showed that the appropriate metaphor is one of an interface between the two fields rather than a common, shared area of inquiry. Thus, the adventures along the frontier cannot be treated as coming from a single community, and the different objectives have to be recognized.

[1] For example, see National Research Council, *Opportunities in Biology*, National Academy Press, Washington, DC, 1989. High-throughput data acquisition is an approach that relies on the large-scale parallel interrogation of many similar biological entities. Such an approach is essential for the conduct of global biological analyses, and it is often the approach of choice for rapid and comprehensive assessment of biological system properties and dynamics. See, for example, T. Ideker, T. Galitski, and L. Hood, "A New Approach to Decoding Life: Systems Biology," *Annual Review of Genomics and Human Genetics* 2:343-372, 2001. A number of the high-throughput data acquisition technologies mentioned in that article are discussed in Chapter 7 of his report.

1.2.1 From the Biology Side

Biologists have a long history of applying tools from other disciplines to provide more powerful methods to address or even solve their research problems. For example, Anton Van Leeuwenhoek's invention of the optical microscope in the late 1600s opened up a previously unknown world and ultimately brought an entirely new vista to biology—namely, the existence of cells and cellular structures. This remarkable revolutionary discovery would have been impossible without the study of optics—and Leeuwenhoek was a clockmaker.

The biological sciences have drawn heavily from chemistry, physics, and more recently, mathematical modeling. Indeed, the reductionist revolution in biological sciences—which led to the current state of understanding of biological function and mechanism at the molecular level or of specific areas such as neurophysiology—in the past five decades began as chemists, physicists, microbiologists, and others interacted and created what is now known as molecular biology. The applications from the physical sciences are already so well established that it is unnecessary to discuss them at length.

Mathematics and statistics have at times played important roles in designing and optimizing biological experiments. For example, statistical analysis of preliminary data can lead to improved data collection and interpretation in subsequent experiments. In many cases, simple mathematical or physical ideas, accompanied by calculations or models, can suggest experiments or lead to new ideas that are not easily identified with biological reasoning alone. An example of this category of contribution is William Harvey's estimation of the volume of the blood and his finding that a closed circulatory system would explain the anomaly in such calculations. Traditionally, biologists have resisted mathematical approaches for various reasons discussed at length in Chapter 10. To some extent, this history is being changed in modern biology, and it is the premise of this report that an acceleration of this change is highly worthwhile.

Approaches borrowed from another discipline may provide perspectives that are unavailable from inside the disciplinary research program itself. In some cases, these lead to a new integrative explanation or to new ways of studying and appreciating the intricacies of biology. In other cases, this borrowing opens an entirely new subfield of biology. The discovery of the helical structure and the "code" of DNA, impossible without crystallography and innovative biological thinking, is one example. The understanding of electrical signaling in neurons by voltage-gated channels, and the Hodgkin-Huxley equations (based on the theory of electrical circuits), constitute another example. Both of these approaches revolutionized the way biology was conducted and required significant, skilled input from other fields.

The most dramatic scenarios arise when major subfields emerge. An example dating back some decades, and described above in another context, is molecular biology, whose tools and techniques (using advanced chemistry, physics, and equipment based on the above) changed the face of biology. A more recent, current example is genomics with its indelible mark on the way that biology as a discipline is conducted and will be conducted for years to come.

The committee believes that from the perspective of the biology researcher, there is both substantial legacy and future promise regarding the application of computing to biological problems. Some of this legacy is manifested in a several-decade development of private-sector databases (mostly those of pharmaceutical companies) and software for data analysis, in public-sector genetic databases, in the use of computer-generated visualization, and in the use of computation to determine the crystal structures of increasingly complex biomolecules.[2]

Several life sciences research fields have begun to take computational approaches. For example, ecology and evolution were among the first subfields of biology to develop advanced computational simulations based on theory and models of ecosystems and evolutionary pathways. Cardiovascular

[2] See, for example, T. Head-Gordon and J.C. Wooley, "Computational Challenges in Structural and Functional Genomics," *IBM Systems Journal* 40(2):265-296, 2001, available at http://www.research.ibm.com/journal/sj/402/headgordon.pdf.

physiology and studies of the structure and function of heart muscle have involved bioengineering models and combined experimental and computational approaches. All of these computational approaches would have been impossible without solid preexisting mathematical models that led to the intuition and formed the basis for the emerging computational aspects.

Nevertheless, genomics research is simply not possible without information technology. It is not an exaggeration to say that it was the sequencing of complete genomes, more than any other research activity, that brought computational and informatics approaches to the forefront of life sciences research, as well as identifying the need for basic underlying algorithms to tackle biological problems. Only through computational analysis have researchers begun to uncover the implications of genomic-scale sequence data. Apart from specific results thereby obtained, such analysis, coupled with the availability of complete genomic sequences, has changed profoundly how many biologists think, conduct research, and plan strategically to address central research problems.

Today, computing is essential to every aspect of molecular and cell biology, as researchers expand their scope of inquiry from gene sequence analysis to broader investigations of biological complexity. This scope includes the structure and function of proteins in the context of metabolic, genetic, and signaling networks, the sheer complexity of which is overwhelming. Future challenges include the integration of organ physiology, catalogs of species-wide phenotypic variations, and understanding of differences in gene expression in various states of health and disease.

1.2.2 From the Computing Side

From the viewpoint of the computer scientist, there is an as-yet-unfulfilled promise that biology may have significant potential to influence computer design, component fabrication, and software. Today, the impact of biology and biological sciences on advances in computing is more speculative than the reverse (as described in Section 1.2.1), because such considerations are, with only a few exceptions, relevant to future outcomes and not to what has been or is already being delivered.

In one sense, this should not be very surprising. Computing is a "science of the artificial,"[3] whereas biology is a science of the natural, and in general, it is much easier for humans to understand both the function and the behavior of a system that they have designed to fulfill a specific purpose than to understand the internal machinery of a biological black box that evolved as a result of forms and pressures that we can only sketchily guess.[4] Thus, paths along which biology may influence computing are less clear than the reverse, and work in this area should be expected to have longer time horizons and to take the form of many largely independent threads, rather than a hierarchy of interrelated or intellectual thrusts.

Nevertheless, exploring why the biological sciences might be relevant to computing is worthwhile in particular because biological systems possess many qualities that would be desirable in the information technology that humans use. For example, computer and information scientists are looking for ways to make computers more adaptive, reliable, "smarter," faster, and resilient. Biological systems excel at finding and learning adequate—but not necessarily optimal—solutions to ill-posed problems on time scales short enough to be useful to them. They efficiently store "data," integrate "hardware" and "software," self-correct, and have many other properties that computing and information science

[3] "We speak of engineering as concerned with 'synthesis,' while science is concerned with 'analysis.' Synthetic or artificial objects—and more specifically prospective artificial objects having desired properties—are the central objective of engineering activity and skill. The engineer, and more generally the designer, is concerned with how things *ought* to be—how they ought to be in order to *attain goals*, and to *function*." H.A. Simon, *Sciences of the Artificial*, 3rd ed., MIT Press, Cambridge, MA, 1996, pp. 4-5.

[4] This is what neuroscientist Valentino Braitenberg called his law of uphill analysis and downhill synthesis, in *Vehicles: Experiments in Synthetic Psychology*, MIT Press/A Bradford Book, Cambridge, MA, 1984. Cited in Daniel C. Dennett, "Cognitive Science as Reverse Engineering: Several Meanings of 'Top-down' and 'Bottom-up'," *Proceedings of the Ninth International Congress of Logic, Methodology and Philosophy of Science*, D. Prawitz, B. Skyrms, and D. Westerstahl, eds., Elsevier Science North-Holland, 1994.

might capture in order to achieve its future goals. Especially for areas in which computer science lacks a well-developed theory or analysis (e.g., the behavior of complex systems or robustness), biology may have the most to contribute.

To hint at some current threads of inquiry, some researchers envision a hybrid device—a biological computer—essentially, an organic tool for accomplishing what is now carried out in silicon. As an information storage and processing medium, DNA itself may someday be the substance of a massively dense memory storage device, although today the difficulties confronting the work in this area are significant. DNA may also be the basis of nanofabrication technologies.

Biomimetic devices are mechanical, electrical, or chemical systems in which an attempt has been made to mimic the way that a biological system solves a particular problem. Successes include robotic locomotion (based on legged movements of arthropods), artificial blood or skin, and others. Approaches with general-purpose applicability are less clearly successes, though they are still intriguing. These include attempts to develop approaches to computer security that are modeled on the mammalian immune system and approaches to programming based on evolutionary concepts.

Hybrid systems are a promising new technology for measurement of or interaction with small biological systems. In this case, hybrid systems refer to silicon chips or other devices designed to interact directly with a biological sample (e.g., record electrical activity in the flight muscles of a moth) or analyze a small biological sample under field conditions. Here the applications of the technology both to basic scientific problems and to industrial and commercially viable products are exciting.

In the domain of algorithms, swarm intelligence (a property of certain systems of nonintelligent, independently acting agents that collectively exhibit intelligent behavior) and neural nets offer approaches to programming that are radically different from many of today's models. Such applications of biological principles to nonbiological computing could have much value, and Chapter 8 addresses in greater detail some possible biological inspirations for computing. Yet it is also possible that a better understanding of information-processing principles in biological systems will lead as well to greater biological insight; so the dividing line between "applying biological principles to information processing" and "understanding biological information processing" is not as clear as it might appear at first glance. Moreover, even if biology ultimately proves unhelpful in providing insight into potential computing solutions, it is still a problem domain par excellence—one that offers interesting intellectual challenges in which progress will require that the state of computing research be stretched immeasurably.

1.2.3 The Role of Organization and Culture

The possibility—or even the fact—that one field may be well positioned to make or facilitate significant intellectual contributions to the other does not, by itself, lead to harmonious interchange between practitioners in the two fields. Cultural and organizational issues are also very much relevant to the success or failure of collaborations across different fields. For example, one important issue is the fact that much of today's biological research is done in individual laboratories, whereas many interesting problems of 21st century biology will require interdisciplinary teams and physical or virtual centers with capable scientists, distributed wherever they work, involved in addressing difficult problems.

Twenty-first century biology will also see the increasing importance of research programs that have a more industrial flavor and involve greater standardization of instruments and procedures. A small example is that reagent kits are becoming more and more popular, as labs realize that the small advantages that might accrue through the use of a set of customized reagents are far outweighed by the savings in effort associated with the use of such kits. A larger example might be shared devices and equipment of larger-scale and assembly-line-like processes that replace the craft work of individual technicians.

As biologists recognize the inherent difficulties posed by the data-intensive nature of these new research strategies, they will require different—and additional—training in quantitative methods and science. Computing is likely to be central, but since the nature and scope of the computing required will go far beyond what is typically taught in an introductory computing course, real advancement of the frontier will require that computer scientists and biologists recognize and engage each other as intellectual coequals. At the same time, computer scientists will have to learn enough about biology to understand the nature of problems interesting to biologists and must refrain from regarding the problem domain as a "mere" application of computing.

The committee believes that such peer-level engagement happens naturally, if slowly. But accelerating the cultural and organizational changes needed remains one of the key challenges facing the communities today and is one that this report addresses. Such considerations are the subject of Chapter 10.

1.3 Imagine What's Next

In the long term, achievements in understanding and harnessing the power of biological systems will open the door to the development of new, potentially far-reaching applications of computing and biology—for example, the capability to use a blood or tissue sample to predict an individual's susceptibility to a large number of afflictions and the ability to monitor disease susceptibility from birth, factoring in genetics and aging, diet, and other environmental factors that influence the body's functions over time and ultimately to treat such ailments.

Likewise, 21st century biology will advance the abilities of scientists to model, before a treatment is prescribed, the likely biological response of an individual with cancer to a proposed chemotherapy regime, including the likelihood of the effectiveness of the treatment and the side effects of the drugs. Indeed, the promise of 21st century biology is nothing less than a system-wide understanding of biological systems both in the aggregate and for individuals. Such understanding could have dramatic effects on health and medicine. For example, detailed computational models of cellular dynamics could lead to mechanism-based target identification and drug discovery for certain diseases such as cancer,[5] to predictions of drug effects in humans that will speed clinical trials,[6] and to a greater understanding of the functional interactions between the key components of cells, organs, and systems, as well as how these interactions change in disease states.[7]

On another scale of knowledge, it may be possible to trace the genetic variability in the world's human populations to a common ancestral set of genes—to discover the origins of the earliest humans, while learning, along the way, about the earliest diseases that arose in humans, and about the biological forces that shape the world's populations. Work toward all of these capabilities has already begun, as biologists and computer scientists compile and consider vast amounts of information about the genetic variability of humans and the role of that variability in relation to evolution, physiological functions, and the onset of disease.

At the frontiers of the interface, remarkable new devices can be pictured that draw on biology for inspiration and insight. It is possible to imagine, for example, a walking machine—an independent set of legs as agile, stable, and energy-efficient as those of humans or animals—able to negotiate unknown terrain and recover from falls, capable of exploring and retrieving materials. Such a machine would overcome the limitations of present-day rovers that cannot do such things. Biologists and computer scientists have begun to examine the locomotion of living creatures from an engineering and biological perspective simultaneously, to understand the physical and biological controls on balance, gait, speed, and energy expended and to translate this information into mechanical prototypes.

[5] J.B. Gibbs, "Mechanism-Based Target Identification and Drug Discovery in Cancer Research," *Science* 287:1969, 2000.

[6] C. Sander, "Genomic Medicine and the Future of Health Care," *Science* 287:1977, 2000.

[7] D. Noble, "Modeling the Heart—From Genes to Cells to the Whole Organ," *Science* 295:1678, 2002.

We can further imagine an extension of present-day bioengineering from mechanical hearts and titanium hip joints to an entirely new level of devices, such as an implantable neural prosthetic that could assist stroke patients in restoring speech or motor control or could enhance an individual's capability to see more clearly in the dark or process complex information quickly under pressure. Such a prosthetic would marry the speed of computing with the brain's capacity for intelligence and would be a powerful tool with many applications.

With the advancement of computational power and other capabilities, there is a great opportunity and challenge in whether human functions can be represented in digital computational forms. One form of representation of a human being is how it is constructed, starting with genes and proteins. Another form of representation is how a human being functions. Human functions can be viewed at many different levels—physioanatomical, motion-mechanical, and psychocognitive, for example. If it were possible to represent a human being at any or all of these functional levels, then a "digital human" could be created inside the computer, to be used for many applications such as medical surgical training, human-centered design of products, and societal simulation. (There are already such simulations at varying levels of fidelity for particular organs such as the heart.)

The potential breadth and depth of the interface of computing and biology are vast. Box 1.1 is a representative list of research areas already being pursued at the interface; Appendix B at the end of this report provides references to more detailed discussions of these efforts. The excitement and challenge of all of these possibilities drive the increasing interest in and enthusiasm for research at the interface.

Box 1.1
Illustrative Research Areas at the Interface of Computer Science and Biology

- Structure determination of biological molecules and complexes
- Simulation of protein folding
- Whole genome sequence assembly
- Whole genome modeling and annotation
- Full genome-genome comparison
- Rapid assessment of polymorphic genetic variations
- Complete construction of orthologous and paralogous groups of genes
- Relating gene sequence to protein structure
- Relating protein structure to function
- In silico drug design
- Mechanistic enzymology
- Cell network analysis-simulation of genetic networks and the sensitivity of these pathways to component stoichiometry and kinetics
- Dynamic simulation of realistic oligomeric systems
- Modeling of cellular processes
- Modeling of physiological systems in health and disease
- Modeling behavior of schools, swarms, and their emergent behavior
- Simulation of membrane structure and dynamic function
- Integration of observations across scales of vastly different dimension and organization for model creation purposes
- Development of bio-inspired autonomous locomotive devices
- Development of biomimetic devices
- Bioengineering prosthetics

1.4 SOME RELEVANT HISTORY IN BUILDING THE INTERFACE

1.4.1 The Human Genome Project

According to Cook-Deegan,[8] the Human Genome Project resulted from the collective impact of three independent public airings of the idea that the human genome should be sequenced. In 1985, Robert Sinsheimer and others convened a group of scientists to discuss the idea.[9] In 1986, Renato Dulbecco noted that sequencing the genome would be an important tool in probing the genetic origins of cancer.[10] Then in 1988, Charles DeLisi developed the idea of sequencing the genome in the context of understanding the biological and genetic effects of ionizing radiation on survivors of the Hiroshima and Nagasaki atomic bombs.[11]

In 1990, the International Human Genome Consortium was launched with the intent to map and sequence the totality of human DNA (the genome).[12] On April 14, 2003, not quite 50 years to the day after James Watson and Francis Crick first published the structure of the DNA double helix,[13] officials announced that the Human Genome Project was finished.[14] After 13 years and $2.7 billion, the international effort had yielded a virtually complete listing of the human genetic code: a sequence some 3 billion base pairs long.[15]

1.4.2 The Computing-to-Biology Interface

For most of the electronic computing age, biological computing applications have been secondary compared to those associated with the physical sciences and the military. However, over the last two decades, use by the biological sciences—in the form of applications related to protein modeling and folding—went from virtually nonexistent to being the largest user of cycles at the National Science Foundation Centers for High Performance Computing by FY 1998. Nor has biological use of computing capability been limited to supercomputing applications—a plethora of biological computing applications have emerged that run on smaller machines.

During the last two decades, federal agencies also held a number of workshops on computational biology and bioinformatics, but until relatively recently, there was no prospect for significant support

[8] Cook-Deegan's perspective on the history of the Human Genome Project can be found in R.M. Cook-Deegan, *The Gene Wars: Science, Politics, and the Human Genome,* W.W. Norton and Company, New York, 1995.

[9] R. Sinsheimer, "The Santa Cruz Workshop," *Genomics* 5(4):954-956, 1989.

[10] R. Dulbecco, "A Turning Point in Cancer Research: Sequencing the Human Genome," *Science* 231(4742):1055-1056, 1986.

[11] C. DeLisi, "The Human Genome Project," *American Scientist* 76:488-493, 1988.

[12] Cook-Deegan identifies three independent public airings of the idea that the human genome should be sequenced, airings that collectively led to the establishment of the HGP. In 1985, Robert Sinsheimer and others convened a group of scientists to discuss the idea. (See R. Sinsheimer, "The Santa Cruz Workshop," *Genomics* 5(4):954-956, 1989.) In 1986, Renato Dulbecco noted that sequencing the genome would be an important tool in probing the genetic origins of cancer. (See R. Dulbecco, "A Turning Point in Cancer Research: Sequencing the Human Genome," *Science* 231(4742):1055-1056, 1986.) In 1988, Charles DeLisi developed the idea of sequencing the genome in the context of understanding the biological and genetic effects of ionizing radiation on survivors of the Hiroshima and Nagasaki atomic bombs. (See C. DeLisi, "The Human Genome Project," *American Scientist* 76:488-493, 1988.) Cook-Deegan's perspective on the history of the Human Genome Project can be found in R. Cook-Deegan, T*he Gene Wars: Science, Politics, and the Human Genome*, W.W. Norton and Company, New York, 1995.

[13] J.D. Watson and F.H. Crick, "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid," *Nature* 171(4356):737-738, 1953.

[14] The "completion" of the project had actually been announced once before, on June 26, 2000, when U.S. President Bill Clinton and British Prime Minister Tony Blair jointly hailed the release of a preliminary, draft version of the sequence with loud media fanfare. However, while that draft sequence was undoubtedly useful, it contained multiple gaps and had an error rate of one mistaken base pair in every 10,000. The much-revised sequence released in 2003 has an error rate of only 1 in 100,000, and gaps in only those very rare segments of the genome that cannot reliably be sequenced with current technology. See http://www.genome.gov/11006929.

[15] Various histories of the Human Genome Project can be found at http://www.ornl.gov/sci/techresources/Human_Genome/project/hgp.shtml.

for academic work at the interface. The Keck Foundation and the Sloan Foundation supported training, and numerous database activities have been supported by federal agencies. As the impact of the Human Genome Project and comparative genomics began to reach the community as a whole, the situation changed. An important step came from the Howard Hughes Medical Institute, which in 1999 held a special competition to select professors in bioinformatics and thus provided a strong endorsement of the role of computing in biology.

In 1999, the National Institutes of Health (NIH) also took a first step toward integrating ad hoc support by requesting an analysis of the opportunities, requirements, and challenges from computing for biomedicine. In June 1999, the Botstein-Smarr Working Group on Biomedical Computing presented a report to the NIH entitled *The Biomedical Information Science and Technology Initiative*.[16] Specifically tasked with investigating the needs of NIH-supported investigators for computing resources, including hardware, software, networking, algorithms, and training, the working group made recommendations for NIH actions to support the needs of NIH-funded investigators for biomedical computing.

That report embraces a vision of computing as the hallmark of tomorrow's biomedicine. To accelerate the transition to this new world of biomedicine, the working group sought to find ways "to discover, encourage, train, and support the new kinds of scientists needed for tomorrow's science." Much of the report focuses on national programs to create "the best opportunities that can be created for doing and learning at the interfaces among biology, mathematics, and computation," and argues that "with such new and innovative programs in place, scientists [would] absorb biomedical computing in due course, while supporting the mission of the NIH." The report also identifies a variety of barriers to the full exploitation of computation for biological needs.

In the intervening 4 years, the validity of the Botstein-Smarr Working Group report vision has not been in question; if anything, the expectations, opportunities, and requirements have grown. Computation in various forms is rapidly penetrating all aspects of life sciences research and practice.

- State-of-the-art radiology (and along with it other fields dependent on imaging—neurology, for example) is highly dependent on information technology: the images are filtered, processed reconstructions that are acquired, stored, and analyzed computationally.
- Genomics and proteomics are completely dependent on computation.
- Integrative biology aimed at predictive modeling is not just computationally enabled—it literally cannot occur in a noncomputational environment.

Biomedical scientists of all stripes are increasingly using public resources and computational tools at high levels of intensity such that very significant fractions of the overall effort are in this domain, and it is highly likely that these trends will continue. Yet many of the barriers to full exploitation of computation in the biological sciences that were identified in the Botstein-Smarr report still remain. One primary focus of the present report is accordingly to consider the intellectual, organizational, and cultural barriers that impede or even prevent the full benefits of computation from being realized for biomedical research.

1.4.3 The Biology-to-Computing Interface

The application of biological ideas to the design of computing systems appears through much of the history of electronic computers, in most cases as an outgrowth of attempts to model or simulate a biological system. In the early 1970s, John H. Holland (the first person in the United States to be awarded a Ph.D. in computer science) pioneered the idea of *genetic algorithms*, which use simulated genetic processes (crossover, mutation, and inversion) to search a large solution space of algorithms.[17]

[16] Available at http://www.nih.gov/about/director/060399.htm.

[17] J.H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, 1975.

This work grew out of research in the 1950s and 1960s to simulate just such processes in the natural world. A second wave of popularity of this technique came after John Koza described genetic programming, which used similar techniques to modify symbolic expressions that comprised entire programs.[18] Both of these approaches are in use today, especially in research and academic settings.

The history of artificial neural networks also shows a strong relationship between attempts to simulate biology and attempts to construct a new software tool. This research predates even the modern electronic digital computers, since Warren McCulloch and Walter Pitts published a model of a neuron that incorporated analog weights into a binary logic scheme in 1943.[19] This was meant to be used as a model of biological neurons, not merely as an abstract computational processing approach. Research on neural nets continued throughout the next decades, focusing on network architectures (particularly random and layered), mechanisms of self-assembly, and pattern recognition and classification. Significant among this research was Rosenblatt's work on perceptrons.[20] However, lack of progress caused a loss of interest in neural networks in the late 1970s and early 1980s. Hopfield revived interest in the field in 1982,[21] and progress throughout the 1980s and 1990s established neural networks as a standard tool for learning and classifying patterns.

A similar pattern characterizes research into cellular automata. John von Neumann's attempts to provide a theory of biological self-assembly inspired him to apply traditional automata theory to a two-dimensional grid;[22] similar work was being done at the same time by Stanislaw Ulam (who may have suggested the approach to von Neumann). Von Neumann also showed that cellular automata could simulate a Turing machine, meaning that they were a system that could provide universal computation. A boom of popularity for cellular automata followed the publication of the details of John Conway's Game of Life.[23] In the early 1980s, Stephen Wolfram made important contributions to formalizing cellular automata, especially in their role in computational theory,[24] and Toffoli and Margolus stressed the general applicability of automata as systems for modeling.[25]

At a more metaphorical level, IBM has taken initiatives in biologically inspired computing. Specifically, IBM launched its Autonomic Computing initiative in 2001. Autonomic computing is inspired by biology in the sense that biological systems—and in particular the autonomic nervous system—are capable of doing many things that would be desirable in complex computing systems. Autonomic computing is conceived as a way to manage increasingly complex and distributed computing environments as traditional approaches to system management reach their limits. IBM takes special note of the fact that "the autonomic nervous system frees our conscious brain from the burden of having to deal with vital but lower-level functions."[26] Autonomic computing, by IBM's definition, requires that a system be able to configure and reconfigure itself under varying and unpredictable conditions, to continually optimize its workings, to recover from routine and extraordinary events that might cause

[18] J.R. Koza, "Genetically Breeding Populations of Computer Programs to Solve Problems in Artificial Intelligence," pp. 819-827 in *Proceedings of the Second International Conference on Tools for Artificial Intelligence*, IEEE Computer Society Press, Los Alamitos, CA, 1990.

[19] W.S. McCulloch and W.H. Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics* 5:115-137, 1943.

[20] R. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Spartan Books, Washington, DC, 1962.

[21] J.J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," *Proceedings of the National Academy of Sciences* (USA) 79(8):2554-2558, 1982.

[22] J. von Neumann, *Theory of Self-reproducing Automata* (edited and completed by A. W. Burks), University of Illinois Press, 1966.

[23] M. Gardner, "MATHEMATICAL GAMES: The Fantastic Combinations of John Conway's New Solitaire Game 'Life'," *Scientific American* 223(October):120-123, 1970.

[24] S. Wolfram, "Computation Theory of Cellular Automata," *Communications in Mathematical Physics* 96:15-57, 1984.

[25] T. Toffoli and N. Margolus, *Cellular Automata Machines: A New Environment for Modeling*, MIT Press, Cambridge, MA, 1987.

[26] G. Ganek and T.A. Corbi, "The Dawning of the Autonomic Computing Era," *IBM Systems Journal* 42(1):5-18, 2003.

some parts to malfunction in a manner analogous to the healing of a biological system, and to protect itself against dangers in its (open) environment.

1.5 BACKGROUND, ORGANIZATION, AND APPROACH OF THIS REPORT

To better understand potential synergies at the BioComp interface and to facilitate the development of collaborations between scientific communities in both fields that can better exploit these synergies, the National Research Council established the Committee on Frontiers at the Interface of Computing and Biology. The committee hopes that this report will be valuable and important to a variety of interested parties and constituencies and that scientists who read it will be attracted by the excitement of research at the interface. To researchers in computer science, the committee hopes to demonstrate that biology represents an enormously rich problem domain in which their skills and talents can be of enormous value in ways that go far beyond their value as technical consultants and also that they may in turn be able to derive inspiration for solving computing problems from biological phenomena and insights. To researchers in the biological sciences, the committee hopes to show that computing and information technology have enormous value in changing the traditional intellectual paradigms of biology and allowing interesting new questions to be posed and answered. To academic administrators, the committee hopes to provide guidance and principles that facilitate the conduct of research and education at the BioComp interface. Finally, to funding agencies and organizations, the committee hopes to provide both a rationale for broadening the kinds of work they support at the BioComp interface and practices that can enhance and create links between computing and biology.

A note on terminology and scope is required for this report. Within the technology domain are a number of interconnecting aspects implied by terms such as computing, computation, modeling, computer science, computer engineering, informatics, information technology, scientific computing, and computational science. Today, there is no one term that defines the breadth of the science and technology within the computing and information sciences and technologies. The intent is to use any of these terms with a broad rather than narrow construction and connotation and to consider the entire domain of inquiry in terms of an interface to life science. For simplicity, this report uses the term "computing" to refer to intellectual domains characterized by roots in the union of the terms above.

Although the words "computing" and "computation" are used throughout this report, biology in the new millennium connects with a number of facets of the exact sciences in a way that cannot be separated from computer science per se. In particular, biology has a synergistic relationship with mathematics, statistics, physics, chemistry, engineering, and theoretical methods—including modeling and analysis as well as computation and simulation. In this relationship, blind computation is no surrogate for insight and understanding. In many cases, the fruits of computation are reaped only after careful and deliberate theoretical analysis, in which the physics, biology, and mathematics underlying a given system are carefully considered. Although much of the focus of this report is on the exchange between biology and computing, the reader should consider how the same ideas may be extended to encompass these other aspects.

Consider, for example, the fact that mathematics plays an essential role in the interpretation of experimental data and in developing algorithms for machine-assisted computing. Computing is implicitly mathematical, and as techniques for mathematical analysis evolve and develop, so will new opportunities for computing.

These points suggest that any specific limits on the range of coverage of this report are artificial and somewhat forced. Yet practicality dictates that some limits be set, and thus the committee leaves systematic coverage of certain important dimensions of the biology-computing interface to other reports. For example, a 2005 report of the Board on Mathematical Sciences (BMS) of the National Research Council (NRC) recommends a mathematical sciences research program that allows biological scientists to make the most effective use of the large amount of existing genomic information and the much larger and more diverse collections of structural and functional genomic information that are being created, covering both current research needs and some higher-risk research that might lead to innovative approaches for the future.[27] The BMS study takes a very broad look at what will be required for bioinformatics, biophysics, pattern matching, and almost anything related to the mathematical foundations of computational biology; thus, it is that BMS report, rather than the present report, that addresses analytical techniques.

Similar comments apply to the present report's coverage of medical devices based on embedded information technologies and medical informatics. Medical devices such as implanted defibrillators rely on real-time analysis of biological data to decide when to deliver a potentially lifesaving shock. Medical informatics can be regarded as computer science applied directly to problems of medicine and health care, focusing on the management of medical information, data, and knowledge for medical problem solving and decision making. Medical devices and medical informatics have many links and similarities to the subject matter of this report, but they, too, are largely outside its scope, although from time to time issues and challenges from the medical area are mentioned. Comprehensive studies describing future needs in medical informatics and medical devices must await future NRC work.

Yet another area of concern unaddressed in this report is the area of ethics associated with the issues discussed here. To ask just a few questions: Who will own DNA data? What individual biomedical data will be collected and retained? What are the ethics involved in using this data? What should individuals be told about their genetic futures? What are the ethical implications of creating new biological organisms or of changing the genetics of already living individuals? All of these questions are important, and philosophers and ethicists have begun to address some of them, but they are outside the scope of this report or the expertise of the committee.

In developing this report, the committee chose to characterize the overarching opportunities at the interface of biology and the computer and information sciences, and to highlight several diverse examples of activities at the interface. These points of intersection broadly represent and illustrate characteristics of research along the interface and include promising areas of exploration, some exciting from a basic science perspective and others from the point of view of novel applications.

Chapter 2 presents perspectives on 21st century biology, a synthesis among a variety of different intellectual approaches to biological research. Chapter 3 is a discussion of the nature of biological data and the requirements that biologists put on data.

Chapter 4 discusses computational tools for biology that help to solve specific and precisely defined problems. Chapter 5 focuses on models and simulations in biology as approaches for exploring and predicting biological phenomena.

Chapter 6 describes the value of a computational and engineering perspective in characterizing biological functionality of interest. Chapter 7 addresses roles in biological research for cyberinfrastructure and technologies for data acquisition.

Chapter 8 describes the potential of computer science applications and processes to utilize biological systems—to emulate, mimic, or otherwise draw inspiration from the organization, behavior, and structure of living things or to make use of the physical substrate of biological material in hybrid systems or other information-processing applications.

Chapter 9 presents a number of illustrative problem domains. These are technical challenges, potential future applications, and specific research questions that exemplify points along the interface of computing and biology. They illustrate the two overarching themes described in Chapter 2, and describe in detail the specific technological goals that must be met in order to successfully meet the challenge.

Chapter 10 is a discussion of the research infrastructure—people and resources need to vitalize the interface. The chapter examines the requisite scientific expertise, the false starts of the past, cultural and other barriers that must be addressed, and the coordinated effort needed to move research at the interface forward.

[27] National Research Council, *Mathematics and 21st Century Biology*, The National Academies Press, Washington, DC, 2005.

Finally, Chapter 11 summarizes key findings about opportunities and barriers to progress at the interface and provides recommendations for priority areas of research, tools, education, and resources that will propel progress at the interface.

Appendix A is a reprint of a chapter from a 1995 NRC report entitled *Calculating the Secrets of Life*. The chapter, "The Secrets of Life: A Mathematician's Introduction to Molecular Biology," is essentially a short primer on the fundamentals of molecular biology for nonbiologists. Appendix B lists some of the research challenges in computational biology discussed in other reports. Short biographies of committee members, staff, and the review coordinator are given in Appendix C.

Throughout this report, examples of relevant work are provided quite liberally where they are relevant to the topic at hand. The reader should note that these examples have generally been selected to illustrate the breadth of the topic in question, rather than to identify the most important areas of activity. That is, the appropriate spirit in which to view these examples is "letting a thousand flowers bloom," rather than one of "finding the prettiest flowers."